Importance sampling in systems simulation

Hopmans, A.C.M.; Kleijnen, J.P.C.

Publication date:
1978

faculteit der economische wetenschappen

RESEARCH MEMORANDUM

[redacted]

TILBURG UNIVERSITY
DEPARTMENT OF ECONOMICS
Postbus 90135 - 5000 LE Tilburg
Netherlands

FEW
73

IMPORTANCE SAMPLING IN SYSTEMS SIMULATION: A PRACTICAL FAILURE?

A.C.M. Hopmans
and
J.P.C. Kleijnen

IMPORTANCE SAMPLING IN SYSTEMS SIMULATION:

A PRACTICAL FAILURE?

A.C.M. Hopmans
(Dr. Neher Laboratory,
P.T.T.,
Leidschendam, the Netherlands)

and

J.P.C. Kleijnen
(Information Systems Group,
Department of Business and Economics,
Katholieke Hogeschool,
Tilburg, the Netherlands)

A study performed under the auspicies of the
Working Group on the Statistical Design and
Analysis of Simulation Experiments,
Section for Operations Research (S.O.R.),
Netherlands Society for Statistics (V.V.S.)

April 1978

Preliminary version.
Comments are sollicited.

CONTENTS

ABSTRACT

A network of servers, known as a grading in telecommunication engineering, is simulated in order to estimate the probability of a customer getting "blocked" (all servers busy). Since blocking is a very rare event (1 ‰ to 5% chance), importance sampling or IS was considered for reduction of the simulation variance. The basic idea of IS is first explained by means of a non-dynamic system. For dynamic systems a method was proposed by Bayes in 1970, which is related to the "virtual measures" published by Carter and Ignall in 1975. For simple queuing systems, we derive the resulting variance, using the renewal (regenerative) property of such systems. For our practical "grading" system several alternative importance regions are investigated. For practical reasons we choose to start an importance region immediately after a call gets blocked (no renewal state). The analysis and simulation experiments for the resulting estimator,yielded the estimated optimal length of the importance region and the optimal number of replications of the region. Unfortunately resulting net variance reduction turned out to be negative.

1. INTRODUCTION

We feel that this paper is unusual in so far as it reports on an unsuccessful research effort. This effort tried to reduce the variability of simulation results by the application of a variance reduction technique (VRT). Not that we think that such unsuccessful investigations have been rare, but reporting such attempts seems to be rare indeed. Nevertheless the documentation of such abortive attempts can be useful: Practitioners may be warned against too optimistic expectations. Theoreticians may be stimulated to revise our approach and, hopefully, come up with a better VRT. Before we explain our particular VRT, we briefly characterize the practical system and its model to which we applied the VRT. Note that a glossary of the major symbols is provided in Appendix 1.

The system of interest is part of a telephone exchange, and is technically known as a grading; for details see Bear (1976). It is convenient to consider this grading as a network of servers. There are g customer generators or traffic sources, and N servers or "lines". In FIG. 1 we see, e.g., that customers (calls) from source 1 have access to server (line) 1, 4, or 5, while line 1 serves customers from the sources 1, 2, or 3 (but only one customer at a time). An equivalent but more customary representation is the diagram of FIG. 2. The actual grading we investigated is more complicated and is shown in FIG. 3. This practical grading shows g=8 customers sources. Each source has k=15 points of entry or "contacts", but since these "servers" are connected to form a common server for several customer sources, only N=45 (not 8 x 15) servers result. When a customer is generated, one of the 15 contacts is selected randomly. If this line is busy, then another line in its "column" is selected cyclically. If all the available 15 (not 45) servers are busy, then the call gets blocked. Ob-

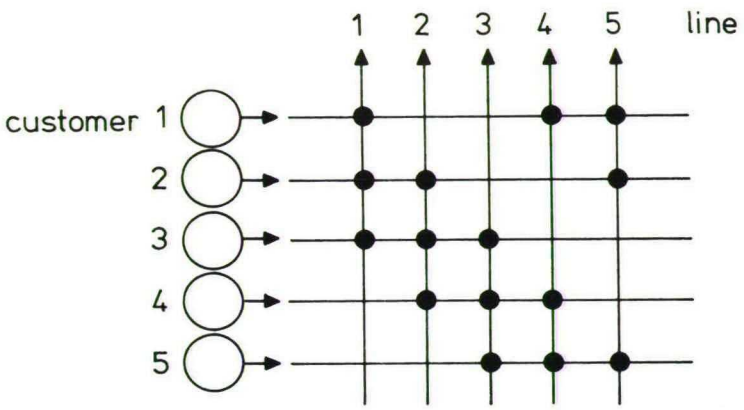

FIG. 1. A simple grading

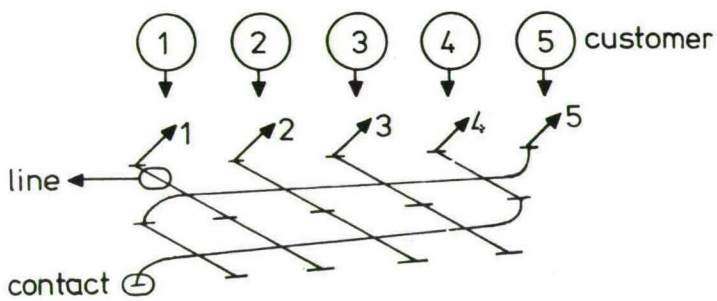

FIG. 2. Equivalent grading as in fig.1.

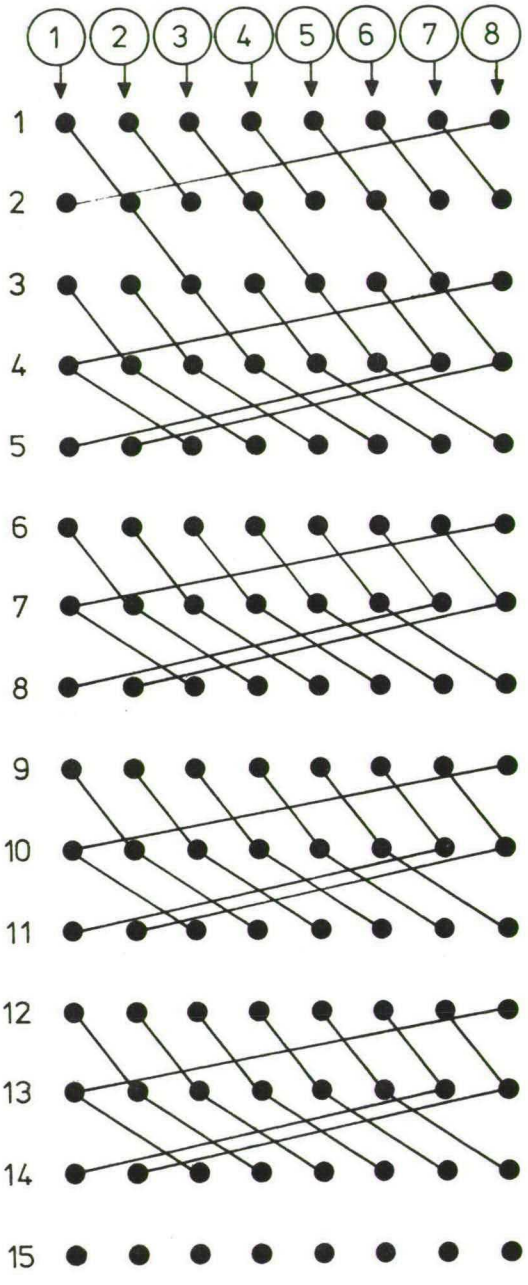

FIG. 3. A realistic grading

serve that unlike traditional queuing systems, the grading type considered here, does not permit customers to wait for service, i.e., if the system is busy new arrivals are lost. (Of course a customer may try again at a later point of time.) We further assume that the 8 customer sources generate demands independently of each other.

How can we estimate the blocking probability? Realistic gradings are too complicated for analytical solutions.[1] Therefore simulation is used. The quantity to be estimated is very small, since a realistic grading is so designed that the blocking probability is between 1 ‰ and 5%. Hence reliable estimates require very long simulation runs, since during long epochs of the simulated history nothing of interest happens, i.e., no blocking occurs!

Elsewhere we discussed how regression analysis could be succesfully combined with simulation by a technique also known as "control variates"; see Hopmans and Kleijnen (1977). Another VRT we applied, is so-called roulette simulation: Because of the Poisson (memoryless) character of our grading, the event-timing administration can be eliminated; for details see de Boer (1969) and Kosten (1948). In this way the required computer time was reduced by a factor 2.

Note that the usual, crude estimator of the steady-state blocking probability p, is

$$\hat{B} = \frac{\text{number of calls lost}}{\text{total number of calls}} \qquad (1.1)$$

The variance of this estimator can be estimated by dividing the total simulation run into a number of subruns (in our case 15 subruns). These subruns can be assumed to give independent blocking probabilities; see Kleijnen (1975, pp. 458-460).

2. IMPORTANCE SAMPLING: NON-DYNAMIC SITUATIONS

The basic idea of importance sampling or IS, was introduced by Kahn and Marshall (1953) as follows: Suppose we wish to estimate ξ, the value of the following integral

$$\xi = \int_{-\infty}^{\infty} g(x)\ f(x)dx = E[g(x)] \tag{2.1}$$

where $f(x)$ is a density function so that (2.1) defines the expected value, denoted by E, of $g(x)$. The crude estimator would sample x from $f(x)$ and compute

$$\hat{\xi} = \frac{1}{n} \sum_{i=1}^{n} g(x_i) \tag{2.2}$$

However, we can also write (2.1) as

$$\xi = \int_{-\infty}^{\infty} \frac{g(x)\ f(x)}{h(x)}\ h(x)\ dx \tag{2.3}$$

So if we choose for $h(x)$ another density function than $f(x)$, then we may sample x from $h(x)$ and compute

$$g^*(x) = \frac{g(x)\ f(x)}{h(x)} = g(x)\ \{\frac{f(x)}{h(x)}\} \tag{2.4}$$

where $f(x)/h(x)$ may be interpreted as a weighing factor. The quantity ξ is estimated by the average of $g^*(x_i)$ $(i=1,\ldots,n)$, analogous to (2.2). It can be derived that the optimal density function $h(x)$ is

$$h_0(x) = \frac{g(x)\ f(x)}{\xi} \tag{2.5}$$

provided $g(x) \geq 0$ for all x. In other words, we sample heavily from the "important" region of x, i.e., from the region where x yields high values for the response $g(x)$, unless the probability of such values is small. Unfortunately, we cannot calculate $h_0(x)$ since it contains the unknown ξ itself!

Nevertheless (2.5) can suggest an adequate approximation to $h_0(x)$. For instance, in Kleijnen (1975, p. 166) the following integral is studied:

$$\xi = \int_v^\infty \left(\frac{1}{x}\right)(\lambda\, e^{-\lambda x})dx \qquad (\lambda, v > 0) \qquad (2.6)$$

so that

$$h_0(x) = \frac{\lambda}{\xi}\frac{1}{x} e^{-\lambda x} \qquad \text{if } x \geq v$$

$$= 0 \qquad \text{if } x < v \qquad (2.7)$$

One possible approximation is to shift the original exponential distribution with parameter λ over a distance v. This reduced the variance drastically: for 4 combinations of λ and v, the variance ranged between 0.7% and 6.5% of the original variance! Kleijnen (1975) gives many more references to importance sampling in non-dynamic situations as in (2.1).

3. IMPORTANCE SAMPLING IN SIMPLE DYNAMIC SYSTEMS

In the simulation of dynamic systems IS is much harder to apply. Various approaches are summarized in Kleijnen (1975, pp. 173-186). Two other studies, however, form the basis for the present study. One approach is that of "virtual measures" introduced by Carter and Ignall (1975). The other approach, closer related to our study, is the method presented by Bayes (1970). The latter approach will be explained in the present section.

Consider a simple queuing system with one server and one customer source. We wish to estimate the probability of a queue q longer than some constant c. This constant c is so high that, hopefully, the above probability is very small. This c may represent a queuing area in a computer

system, a doctor's office, etc. The crude estimator is

$$\hat{P}(q \geq c) = \frac{\text{total time during which } q \geq c}{\text{total simulated time}} \qquad (3.1)$$

Obviously the "rare event" $q \geq c$ tends to happen more frequently when the simulation enters a "heavy loaded" period of the system. In other words if, say, $c = 15$ then $q \geq 15$ is expected to occur more frequently when a customer enters a system with, say, $q = 9$ customers waiting (so that q jumps to 10). Bayes (1970) proposes to repeat that part of the simulation run which started from such a situation; see the dotted lines[2] in FIG. 4. He further proposes to stop such a "replication" (dotted line) as soon as the queue drops below $q = 10$.

Obviously we have to correct for the fact that the important regions are sampled more frequently. Therefore we take the averages of the times during which the rare event occured τ, and the lengths of the important regions t; the time it takes before the critical region is reached is denoted by θ; see FIG. 5.[3] Summarizing, we wish to estimate

$$p \equiv P(q \geq 15) = \frac{E(\tau)}{E(\theta+t)} \qquad (3.2)$$

The crude estimator (3.1) can be written in the symbols of FIG. 5 as [4]

$$\hat{P}_c \equiv \hat{P}(q \geq 15 | \text{crude}) = \frac{\Sigma \tau_i}{\Sigma(\theta_i + t_i)} = \frac{\bar{\tau}}{(\bar{\theta}+\bar{t})} \qquad (3.3)$$

The IS estimator with m replications ($m>1$) - see FIG. 6 - is analogous to (3.3):

$$\hat{P}_{IS} \equiv \hat{P}(q \geq 15 | IS) = \frac{\Sigma \bar{\tau}_i}{\Sigma(\theta_i + \bar{t}_i)} \qquad (3.4)$$

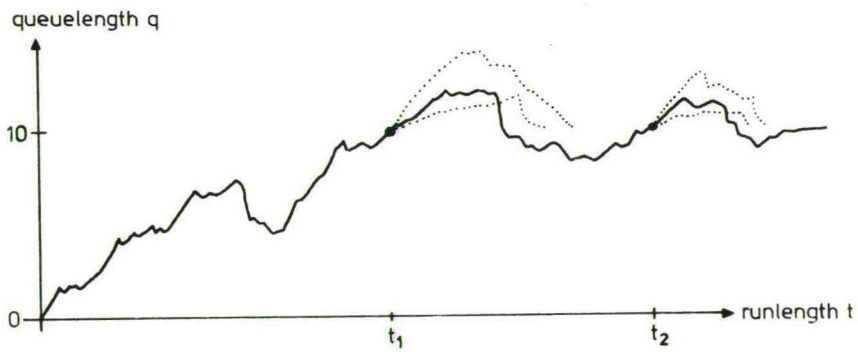

FIG. 4. Replicating important regions ($q \geq 10$)

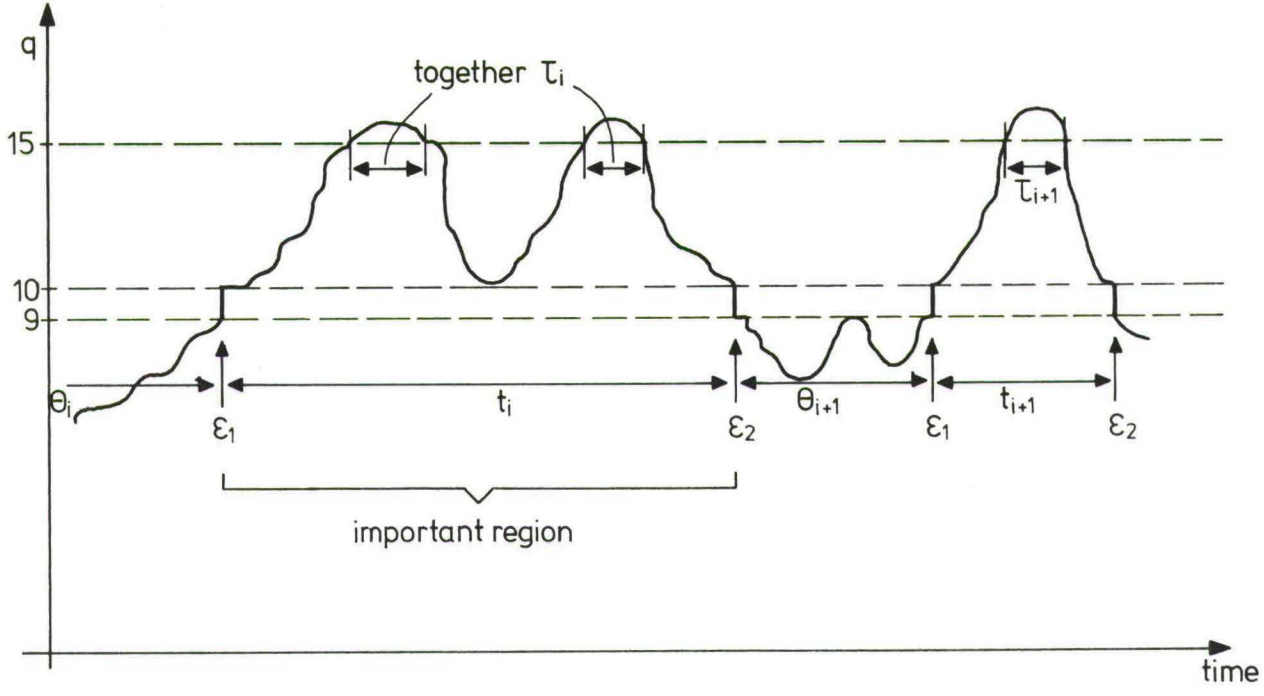

FIG. 5. Importance sampling symbols θ, t and τ

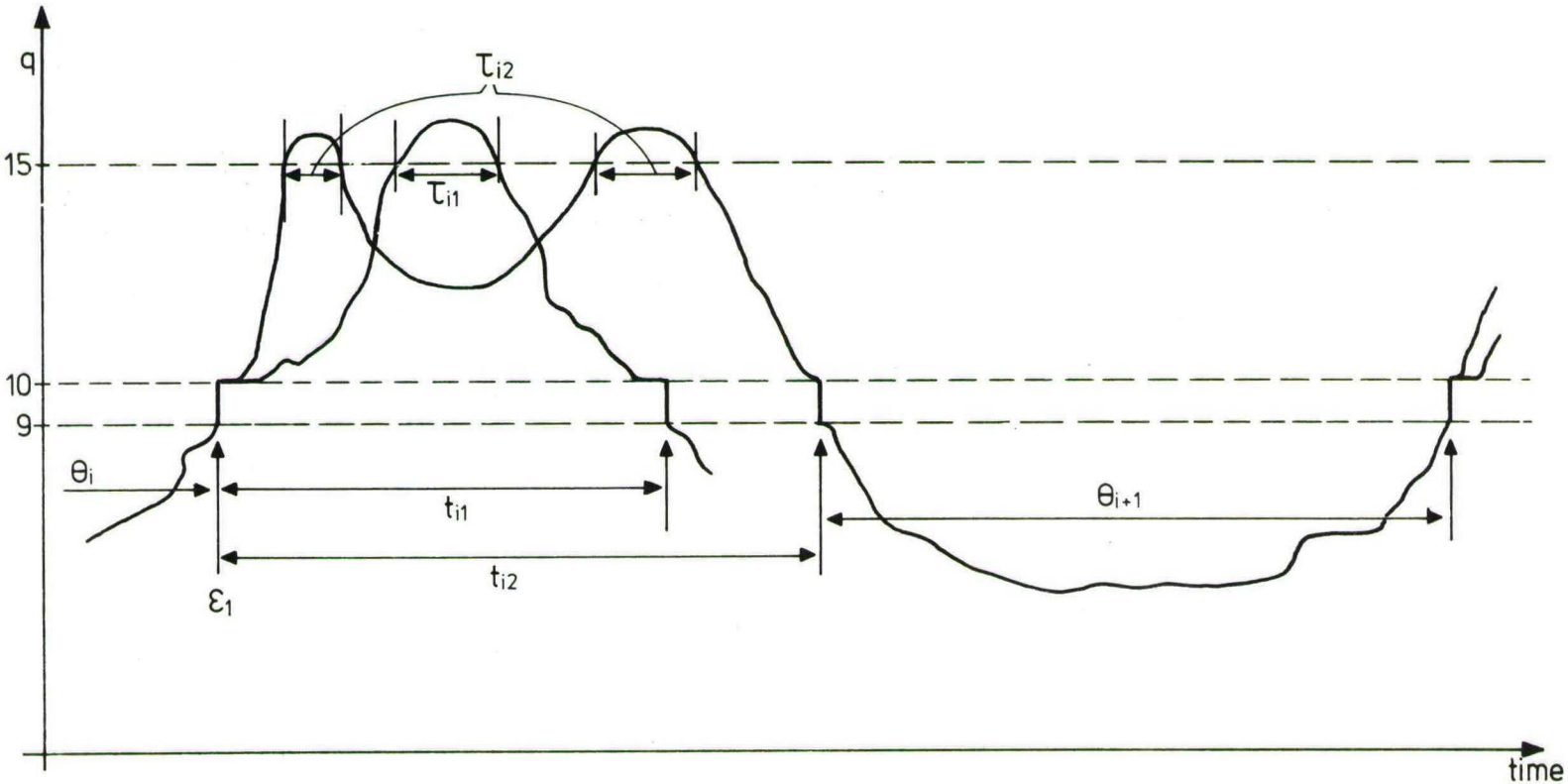

FIG. 6. Replicated important regions (m=2)

with the averages per replicated important region i :

$$\bar{\tau}_i = \sum_{j=1}^{m} \tau_{ij}/m, \qquad (3.5)$$

and

$$\bar{t}_i = \sum_{j=1}^{m} t_{ij}/m \qquad (3.6)$$

For a mathematical derivation of the "obvious" estimator (3.4) we refer to Appendix 2. Note that the original "background" simulation run is formed by continuing one arbitrary replication. It is convenient (and unbiased) to continue the last replication.

The variance of both the crude and the IS estimators in this simple simulation model, can be derived analytically, using the "renewal" or regenerative property. Phrased informally, once we know that q jumped from 9 to 10 (event ε_1) we know enough to continue the simulation run; remember that the assumed Poisson arrival and service processes imply a memoryless system. Hence all cycles (epochs, tours) starting in ε_1 are identically and independently distributed; an alternative renewal point is ε_2. For a discussion of the renewal property in a simulation context we refer to Iglehart (1975). Applying the regenerative property we prove in Appendix 3 that

$$\text{var } (\hat{P}_c) = \frac{\sigma_\tau^2 - 2p \text{ cov}(\tau,t) + p^2 \sigma_{t+\theta}^2}{n.[E(t+\theta)]^2}$$

$$\text{for } n \to \infty \qquad (3.7)$$

Using a similar derivation for var $(\hat{P}_{IS})$ we find that the gross variance reduction is

$$VR_{gross} = m/\{1 + (m-1)\ p^2\ \sigma_\theta^2/\sigma_z^2\ \} \qquad (n\to\infty) \qquad (3.8)$$

where σ_z^2 is a shorthand notation for the numerator of (3.7). The net variance reduction corrects for the (m-1) extra subruns of length t with

$$E(t) = P(q \geq 10)\ .\ E(T) \qquad (3.9)$$

where $T = \theta + t$. Hence the extra simulation length with which to correct the gross variance reduction, yields the factor

$$\frac{(m-1).P(q\geq 10).E(T)+E(T)}{E(T)} = (m-1).P(q\geq 10)+1 \qquad (3.10)$$

so that the net variance reduction follows from (3.8) and (3.10):

$$VR_{net} = m/\{1+(m-1)P(q\geq 10)\}\{1+(m-1)P^2(q\geq 15)\frac{\sigma_\theta^2}{\sigma_z^2}\} \qquad (3.11)$$

It is not obvious in which direction VR_{net} reacts to changes in the start of the importance region (e.g. starting from 12 instead of 10), and changes in the probability of the rare event (e.g. defining the rare event as $q \geq 20$ instead of $q \geq 15$). We have not investigated this problem. However, in the next sections we do investigate a similar selection problem for our more complicated practical system.

4. IMPORTANCE SAMPLING IN A PRACTICAL "GRADING" SYSTEM

Let us introduce the following terminology: The "importance boundary" denotes the start of the importance region in which $m \geq 1$ replications are simulated. If $m = 1$ then IS "degenerates" into crude sampling. In crude simulation the estimator of the steady-state blocking probability B is the average of M subrun probabilities $\hat{B}$:

$$\bar{B} = \sum_{k=1}^{M} \hat{B}_k \; / \; M \qquad (4.1)$$

with subrun probability estimator

$$\hat{B}_k = L_k \; / \; SS_k \qquad (4.2)$$

where

L_k: number of calls blocked or "lost" in subrun k,

SS_k: total number of calls in subrun k (sample size). Since SS_k is kept constant in all subruns, we may drop the index k, i.e., $SS_k = SS = 10{,}000$.

The application of IS to the grading of FIG. 3 becomes troublesome because of the complexity of this system. In the preceding section a renewal state (completely specifying the system's state) was the value of the queuelength q (Poisson, memoryless arrival and service processes were assumed). In theory, assuming Poisson processes for the grading, a possible renewal state could be defined by specifying for each individual line whether this "server" is busy or idle. However, there are as many as N=45 lines so that a return to this specific system state will take very long, as the total number of possible states is $2^{45} \approx (3.5)(10^{13})$. A renewal state does not necessarily yield a good starting point for an importance region. FIG. 3 shows that 15 busy lines can already block some customer source. The other extreme, all N=45 lines busy, would imply that all customer sources are blocked. Normally calls get blocked before this extreme is reached.

An alternative starting point for the importance region is provided by the total number of busy lines,or TBL. Compared to the above renewal states, we ignore the identity of the lines. The arrows in FIG. 7 show that the importance region starts as soon as we cross the boundary line from

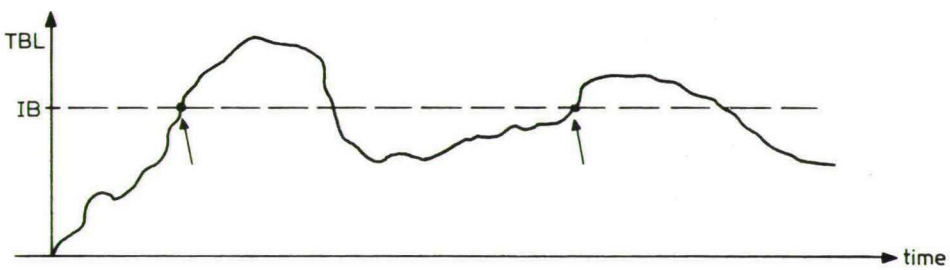

FIG. 7. Total number of busy lines (TBL) as importance boundary

below. This starting point is not a renewal state! To continue the simulation we would have to know not only the TBL value but also the identity of the busy lines. Even with Poisson processes, a different assignment of identities would result in a different subsequent history. Therefore it does not make much sense to decide to end the importance replication when the line in FIG. 7 returns to the boundary line from above, a procedure described in the preceding section and originally proposed by Bayes (1970).

Selecting TBL as a boundary condition does not provide a renewal state but does yield a more frequently occurring system state from which to start replications. Since it is no renewal state the length of the replication SS is made constant, instead of being dependent on the return to the same TBL value from "above". The length of a replication is defined by the total number of generated calls. Replications starting from the same boundary point are made independent by the use of different random number streams.[5)]

Other boundary conditions may be considered. We restricted our study to the following options:
(1) The total number of busy lines TBL; see above.
(2) If all 15 lines serving one specific customer source are busy, and this particular source generates a call, then this call gets blocked. Therefore we start an importance region as soon as any customer source shows 15 busy servers.
(3) Immediately after a call gets blocked, an importance region is started.
In pilot studies we found that the first two options do not lead to importance regions in which many more calls get blocked than in the other regions (called θ in FIG. 6). Therefore we shall concentrate in this paper on the results with the "more promising" option 3.

5. RESULTS FOR A SPECIFIC IMPORTANCE BOUNDARY

As we mentioned in the preceding section we conjectured that an important region starts as soon as a call gets blocked. In other words we expect that lost calls are clustered. This conjecture is checked by performing a pilot simulation run, and measuring the number of calls between two consecutive blocked calls: "interarrival time" of blocked calls, or IA. The resulting frequency diagram is shown in FIG. 8 with double logarithmic scaling.[6] This figure suggests that a good approximation is

$$P(IA = k) = 0.175\ k^{-1.06} \qquad (k=2,3,\ldots,512) \qquad (5.1)$$

The mean and median are 67.8 and 12 respectively, i.e., the distribution is very asymmetric and suggests that the "rare events" (lost calls) occur in clusters. This result seems an encouraging indication of a useful importance boundary definition! The length of replication j (j=1,...,m) in the important region is denoted by a constant SSR. Within subrun k (k=1,...,15) the important region may be entered again later on; see index i below (i=1,...,η). When the important region is entered, the system state is saved. Hence we may imagine that after the whole run has terminated, (m-1) replications (of length SSR) are performed starting from the boundary state i within subrun k. These (m-1) extra histories are simulated using a separate random number stream. See also FIG. 9 where only one subrun is pictured, and the index k is deleted in the symbols.[7]

Consider a subrun k (without importance region replications, i.e., m=1). Let η_k denote the number of times an important region is entered within subrun k. Hence η_k calls are blocked in subrun k outside the importance regions. Each importance region is replicated (m-1) extra times, and has length SSR. The number of calls lost within an importance

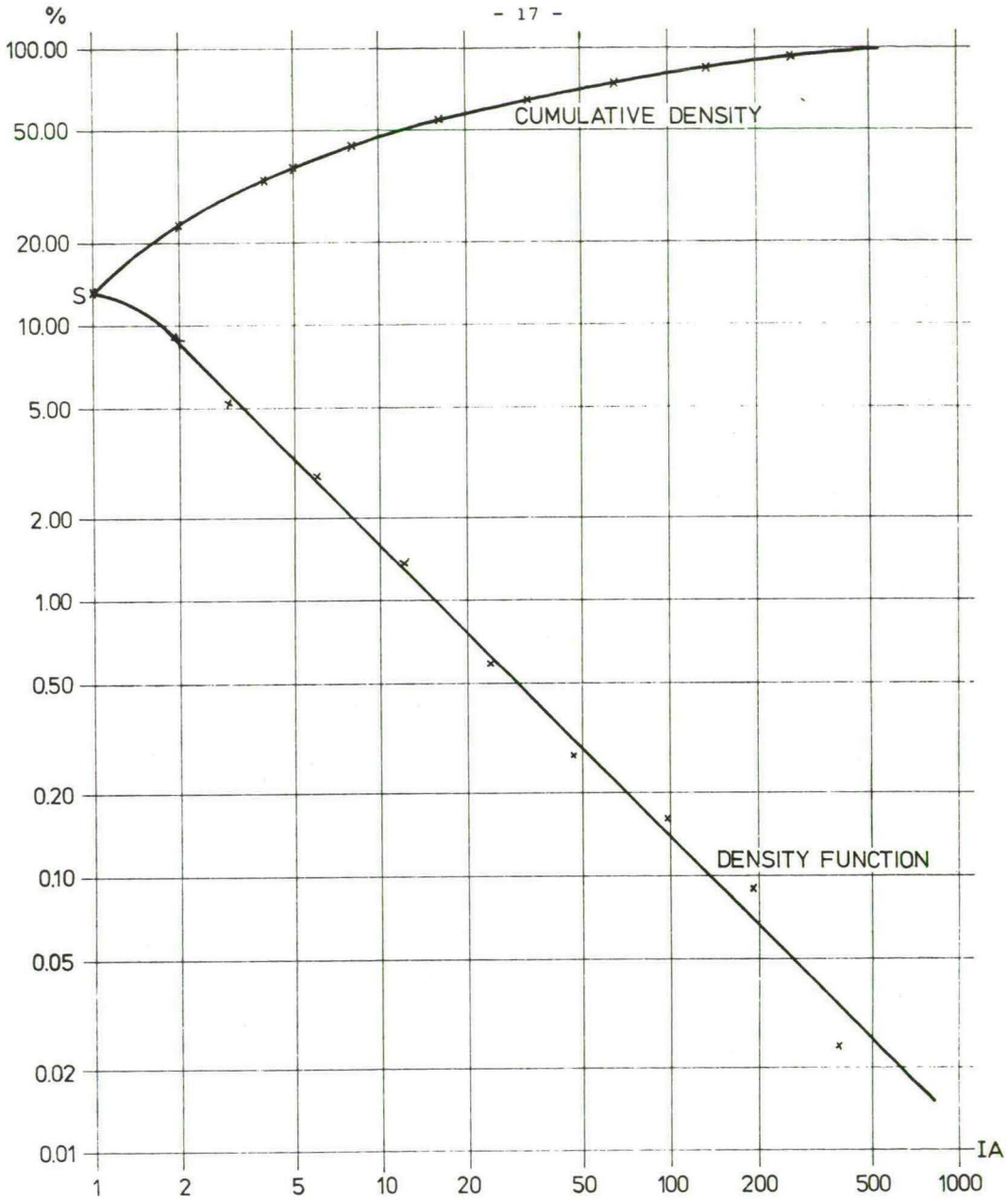

FIG. 8. Number of calls between 2 consecutive lost calls

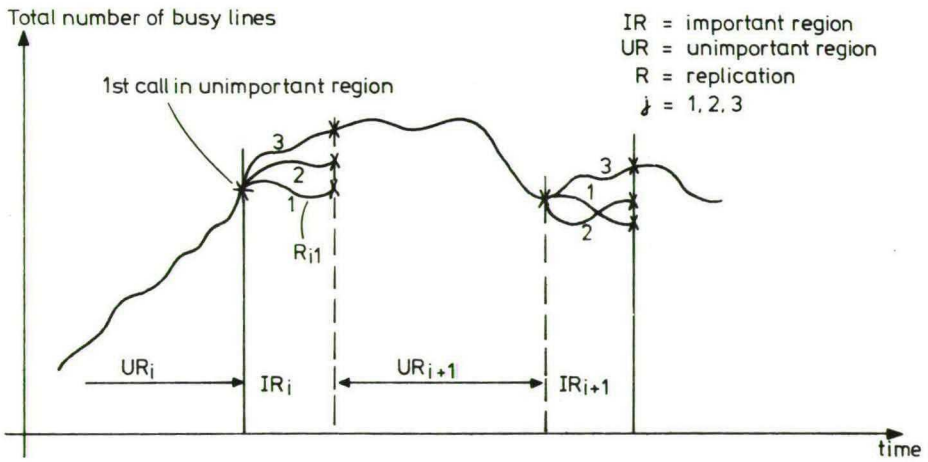

FIG. 9. An importance sampling realization

replication is denoted by LR. Since a subrun has total length SS, the estimated blocking probability is

$$\hat{B}_k = \{\eta_k + \sum_{i=1}^{\eta_k} (\sum_{j=1}^{m} LR_{kij}/m)\}/SS \qquad (5.2)$$

In Appendix 4 we prove that this estimator is an unbiased estimator of the steady-state blocking probability .

To derive the variance of the importance sampling estimator (5.2), we denote the numerator by L_k; the denominator is a constant. Since L_k comprises a summation over a stochastic number of terms - η_k in the first Σ sign - we use the well-known formula (see Keeping, 1962, p. 398):

$$\text{var}(y) = \underset{x}{E}\{\text{var}(y|x)\} + \underset{x}{\text{var}}\{E(y|x)\} \qquad (5.3)$$

In Appendix 5 we derive that

$$\text{var}(\hat{B}_k) = [\{1 + E(LR)\}^2 \text{var}(\eta) + E(\eta)\,\text{var}(LR)/m]/SS^2 \qquad (5.4)$$

In that appendix we assume that several variables are independent, an assumption that seems realistic. If the assumption, however, is violated, then positive correlation may be expected, so that (5.4) needs the addition of some positive terms. Hence (5.4) is a lower bound, so that the derived variance reduction is an upper bound.

The variance without importance sampling follows from (5.4) by substituting m=1. Hence the gross variance reduction (neglecting repeated sampling effort) is:

$$VR_{gross} = \frac{\{1 + E(LR)\}^2 \, var(\eta) + E(\eta) \, var(LR)}{\{1 + E(LR)\}^2 \, var(\eta) + \frac{1}{m} E(\eta) \, var(LR)}$$

$$= \frac{f_1 + f_2}{f_1 + \frac{1}{m} f_2} = \frac{c}{f_1 + \frac{1}{m} f_2} \qquad (5.5)$$

where the terms f_1 and f_2 are introduced to simplify the following presentation. The effect of repetitions in the importance region is shown by the factor 1/m in (5.5). Those replications have more effect as the magnitude of f_2 is large relative to f_1. Obviously, the sum $f_1 + f_2$, i.e., the numerator in (5.5), is independent of the partitioning of the total simulation run into "important" and "unimportant" regions. The shares of f_1 and f_2 in the constant c, depend (among other things) on the length of the importance replication SSR. It is interesting to consider two limiting cases:

Case 1: SSR = 0.

Since there are no replications in the important region, LR = 0. Hence

$$c = f_1 + f_2 = \{1 + 0\}^2 \, var(\eta) + E(\eta).0 = var(\eta) \qquad (5.6)$$

Case 2: SSR approaches SS.

A subrun starts in an unimportant region. As soon as a call gets blocked, the rest of the subrun is replicated as an important region. (So SS-1 is a weak upperbound for SSR.) Consequently $\eta_k = 1$ and

$$c = f_1 + f_2 = \{1 + E(LR)\}^2.0 + 1.var(LR)$$

$$= var(LR) \qquad (5.7)$$

Comparing cases 1 and 2, we see that f_2 is maximal relative to f_1, if SSR approaches SS. Considering (5.5) this means

that in that case the effect of replications in the important region is maximal. So we might jump at the conclusion that the length of the importance replication should be as long as possible. However, as the replication moves on, the effect of its starting point diminishes! What is the net effect of these two conflicting reasonings? We shall present numerical results below.

The gross variance reduction (5.5) needs correction for the extra sampling effort ESE, with expected value

$$E(ESE) = E(\eta)(m-1)\ SSR \tag{5.8}$$

Hence the net variance reduction is

$$VR_{net} = \frac{(f_1 + f_2)}{(f_1 + \frac{1}{m} f_2)} \quad \frac{SS}{\{SS + E(\eta)(m-1)SSR\}} \tag{5.9}$$

where f_1 and f_2 both depend on SSR. To maximize (5.9) we need to select optimal values for SSR, length of replication, and m, number of replications. Note that the factor m is not involved in any of the other factors in (5.9). VR_{net} is maximal if its denominator is minimal. Hence we determine the partial derivative[8] $\partial/\partial m$ and solve $\partial/\partial m = 0$. The optimal number of replications is found to be

$$m_\sigma = [\frac{f_2}{f_1} \quad \frac{\{SS - E(\eta).SSR\}}{E(\eta).SSR}]^{\frac{1}{2}} \tag{5.10}$$

The functions $f_1(SSR)$ and $f_2(SSR)$ are not explicitly known, so that we cannot compute the optimal value m_0 from (5.10). Neither can we compute the optimal SSR as $\partial(VR)/\partial(SSR)$ cannot be made explicit. Therefore we estimate f_1 and f_2 besides $E(\eta)$, for various SSR values, using a pilot simulation run. In this simulation run m = 1, i.e., no importance sampling is needed! This results in Table 1 which we can explain

as follows.

(i) As columns 2 and 3 show, when a greater part (SSR) of the total subrun is considered as forming an importance region, then the remaining number of blocked calls (new entries of an importance region) necessarily decreases, i.e., $\bar{\eta}$ decreases where $\bar{\eta}$ is the average of 15 subruns.

EXP. (1)	SSR (2)	$\bar{\eta}$ (3)	$\hat{f}_1$(SSR) (4)	$\hat{f}_2$(SSR) (5)	$\hat{m}_0$ (6)	$\hat{VR}_{net}$ (7)
1	0					
2	1	137.4	426.86	16.85	1.68	1.006
3	5	105.6	406.88	57.32	1.59	1.016
4	10	89.8	321.43	97.08	1.75	1.040
5	20	74.7	318.78	146.17	1.62	1.040
6	25	69.5	257.53	166.39	1.75	1.063
7	30	65.0	271.71	195.59	1.72	1.063
8	50	56.4	292.28	241.18	1.45	1.032
9	75	48.5	313.36	299.81	1.29	1.016
10	100	42.9	238.82	309.02	1.31	1.019

Table 1: Estimated m_0 and VR_{net}

(ii) Columns 4 and 5 are the estimates of f_1 and f_2 defined in (5.5). Note that each subrun yields several importance regions (namely η_k); each region results in a single value for lost calls LR.[9)]

(iii) Substituting the values $\hat{f}_1$ and $\hat{f}_2$ into (5.10) yields $\hat{m}_0$ in column 6, and the estimated variance reduction based on (5.9), in column 7.

Our conclusions based on Table 1 are:

(i) The maximal net variance reduction is less than 6.3%.[10)]

(ii) The corresponding optimal number of replications is, after rounding to the nearest integer, only 2. For too long importance replications (SSR $\geq$ 50) this number is just 1, i.e., no importance sampling should be done!
(iii) The optimum length of the importance region SSR is about 25.

The above conclusions are based on estimates only (but the numbers in columns 6 and 7 do not show wild oscillations). We checked our conclusions by actually executing an importance sampling experiment with SSR = 25 and m = 2. The gros variable reduction was 1.0085 but the net variance reduction (accounting for the one extra run per importance region) was only 0.859. In other words,the variance even increased with 14%.(Of course these estimates could be inaccurate again.)

6. CONCLUSION

Importance sampling was originally developed for the evaluation of integrals such as (2.1). In that area dramatic variance reductions have been realized (e.g. a factor 100). The extension of this technique to dynamic systems was tried by several authors. The variant that inspired our study was developed by Bayes (1970) and shows some relationships with the "virtual measures" of Carter and Ignall (1975). However, we applied Bayes' procedure to a much more complex system, namely a server network or "grading" occurring in telephone exchanges. In such a grading renewal states could be detected but they could not be utilized since the return to such a state takes too long for practical purposes, and does not necessarily start an important region.

The crucial issue is to define situations (states) which initialize an "important region", i.e., a part of the simulation run in which many important - but rare - events

are expected to occur. Three alternative "importance boundaries" were investigated. This report concentrated on the boundary that seemed most promising, namely, an important region starts immediately after a customer (call) gets blocked, for we found that lost calls tend to occur in clusters.

Next we were confronted with two tactical questions: how long to continue sampling in the importance region, and how often to repeat this sampling? We derived a formula for the (net) variance reduction (correcting for the additional sampling effort). This formula could not be solved analytically for the optimal sampling length and replication number. Therefore estimates were substituted based on a pilot simulation run. The results indicate that the way we applied importance sampling in our particular system, resulted in a net variance increase!

The lesson for practitioners may be not to use importance sampling since the resulting variance reduction may very well be poor. Moreover, its application is not so straightforward as that of some other variance reduction techniques. (Nevertheless a side-benefit was that during our analysis we gained an improved understanding of the way our system behaves as a stochastic process.) Our study may be of interest to theoreticians, in so far as it provides a challenge to improve our importance sampling technique which seems of particular value in systems characterized by "rare" events.

REFERENCES

BAYES, A.J., Statistical techniques for simulation models. AUSTRALIAN COMPUTER JOURNAL, 2, no. 4, Nov. 1970, pp. 180-184.

BEAR, D., PRINCIPLES OF COMMUNICATION-TRAFFIC ENGINEERING. P. Peregrinus Ltd., Southgate House, Stevenage (England), 1976.

CARTER, G., and E. IGNALL, Virtual measures: a variance reduction technique for simulation. MANAGEMENT SCIENCE, 21, no. 6, Feb. 1975, pp. 607-616.

DE BOER, J., Toepassing van Kosten's lotingsas in simulaties. (Applying Kosten's sampling axis to simulation.) STATISTICA NEERLANDICA, 23, no. 3, 1969, pp. 243-248.

HOPMANS, A.C.M. and J.P.C. KLEIJNEN, REGRESSION ESTIMATORS IN SIMULATION. Report FEW-70, Department of Economics, Katholieke Hogeschool, Tilburg (Netherlands), Dec. 1977.

IGLEHART, D.L., SIMULATING STABLE STOCHASTIC SYSTEMS, V: COMPARISON OF RATIO ESTIMATORS. Technical report no. 86-14, Control Analysis Corporation,Palo Alto (California), July 1974.

IGLEHART, D.L., STATISTICAL ANALYSIS OF SIMULATIONS. Technical Report no. 86-18, Control Analysis Corporation, Palo Alto, California, July 1975.

KAHN, H. and A.W. MARSHALL, Methods of reducing sample size in Monte Carlo computations. JOURNAL OPERATIONS RESEARCH SOCIETY OF AMERICA, 1, no. 5, Nov. 1953, pp. 263-278.

KEEPING, E.S., INTRODUCTION TO STATISTICAL INFERENCE. D. Van Nostrand Company, Inc., Princeton, 1962.

KOSTEN, L., On the measurement of congestion quantities by means of fictitious traffic. HET P.T.T. BEDRIJF, 2, 1948-1949, pp. 15-25.

ACKNOWLEDGEMENT

This research was done by the authors as members of the Working Group on the Statistical Design and Analysis of Simulation Experiments, chaired by J.P.C. Kleijnen, under the auspicies of the Section for Operations Research (SOR) of the Netherlands Society for Statistics (VVS). Many critical questions and helpful comments were received from the members of the Working Group, especially B. Sanders and R. van der Ven (P.T.T., The Hague), G. Horstmeier and R. Sierenberg (Delft University), T. Boulogne and R. van der Ham (ECT, Rotterdam).

NOTES

1) We assume identical Poisson customer sources and exponential service times. Then a Markov process results. This system would require the solution of 2^N equations ($2^N \approx 3.5 \times 10^{13}$).

2) Our terminology is such that m "replications" means that 1 "replication" is part of the background or base run, and (m-1) "replications" are duplicates.

3) Observe that τ_{ij} may consist of non-consecutive epochs during which $q \geq 15$, within the j th replication. Further t is the time between the events ε_1, and the next event ε_2, and θ is the time between ε_2 and a next event ε_1.

4) Note that $E(\tau)/E(\theta+t) \neq E[\tau/(\theta+t)]$, so that (3.3)is a biased estimator.Asymptotically this estimator becomes unbiased. Alternative "ratio" estimators are surveyed in Iglehart (1974). However, in crude estimation it is possible to fix the total simulation runlength so that the denominator of (3.3) becomes deterministic.

5) Replications starting at a "later" boundary point (say, the righthand arrow in FIG. 7), are theoretically dependent on the previous history, and hence on the last replication of the preceding importance region. If importance regions are "far" apart, this dependence may be ignored for practical purposes.

6) FIG. 8 shows that all observations are close to a linear line, with the exception of the starting point, denoted by S.

7) For completeness sake we mention that the simulation is started in the empty state (all lines free), and the total run is cut into 15 subruns, each comprising 10,000 calls. No subrun starts in an important region.

8) $\frac{\partial(\text{denominator})}{\partial m} = f_2.E(\eta).SSR.m^{-2} + E(\eta).SSR.f_1 - SS.f_2.m^{-2}$

It is easy to check that (5.10) defines a minimum indeed, and not a maximum or saddlepoint.

9) We compute

$$\overline{\overline{LR}}... = \frac{1}{15} \sum_{k=1}^{15} . \frac{1}{\eta_k} \sum_{i=1}^{\eta_k} LR_{ki}$$

and

$$S^2_{LR_{ki}} = \frac{1}{\eta_k - 1} \{ \Sigma LR^2_{ki1} - (\Sigma LR_{ki1})^2 / \eta_k \}$$

so that

$$\overline{S^2_{LR_{ki}.}} = \frac{1}{15} \sum_{k=1}^{15} S^2_{LR_{ki}.}$$

Then

$$\hat{f}^1 = (1 + \bar{\bar{L}}R...)^2 \, s_\eta^2$$

and

$$\hat{f}_2 = \bar{\eta} \cdot s^2_{\bar{L}R_{ki.}}$$

10) Remember that below (5.4) we noted that if actually some variables are dependent then our formula gives an upper bound for the variance reduction, so that this 6.3% is an estimated upper bound.

APPENDIX 1: GLOSSARY OF MAJOR SYMBOLS

N	: total number of servers in the grading
k(1,2,...,M)	: subrun index
i(1,2,...,η_k)	: important region index within a subrun
j(1,2,...,m)	: replication index within an important region
M = 15	: number of subruns in a simulation run
η_k	: number of important regions in subrun k
m	: number of replicated simulations in <u>one</u> important region
SS_k = SS	: <u>S</u>ample <u>S</u>ize = number of generated calls in a subrun
SSR	: <u>S</u>ample <u>S</u>ize <u>R</u>eplication = numbers of generated calls in a replication
UR	: an <u>U</u>nimportant <u>R</u>egion in a subrun
IR	: an <u>I</u>mportant <u>R</u>egion in a subrun
TBL	: <u>T</u>otal number of <u>B</u>usy <u>L</u>ines
L_k	: number of lost calls in subrun k
$LR_{ki\,j}$	: number of lost calls in the j-th replicated simulation in the i-th important region of subrun k
B = P(b)	: call-blocking probability

$P(b|UR)$: call-blocking probability in an UR

$P(b|IR)$: call-blocking probability in an IR

APPENDIX 2: DERIVATION OF IS ESTIMATOR (3.4)

Obviously

$$P(q \geq 15) = P(q \geq 15 | q \geq IB)\ P(q \geq IB). \tag{A2.1}$$

An estimator for the conditional probability is

$$\hat{P}(q \geq 15 | q \geq IB) = \sum_{i=1}^{\eta} \sum_{j=1}^{m} w_{ij} \cdot \frac{\tau_{ij}}{t_{ij}} \tag{A2.2}$$

where η = number of "important regions" in the run

$$w_{ij} = \text{weighing factor} = \frac{t_{ij}}{t..} \tag{A2.3}$$

$$t.. = \sum_{i=1}^{\eta} \sum_{j=1}^{m} t_{ij} \tag{A2.4}$$

Defining

$$\bar{\tau}_{i\cdot} = \frac{1}{m} \sum_{j=1}^{m} \tau_{ij} \tag{A2.5}$$

and

$$\bar{t}_{i\cdot} = \frac{1}{m} \sum_{j=1}^{m} t_{ij} \tag{A2.6}$$

(A2.2) becomes

$$\hat{P}(q \geq 15 | q \geq IB) = \frac{m \sum_{i=1}^{\eta} \bar{\tau}_{i\cdot}}{m \sum_{i=1}^{\eta} \bar{t}_{i\cdot}} = \frac{\sum_{i=1}^{\eta} \bar{\tau}_{i\cdot}}{\sum_{i=1}^{\eta} \bar{t}_{i\cdot}} \tag{A2.7}$$

An estimator for $P(q \geq IB)$ is

$$\hat{P}(q \geq IB) = \sum_{i=1}^{\eta} \cdot g_i \cdot \frac{\bar{t}_{i.}}{\theta_i + \bar{t}_{i.}} \tag{A2.8}$$

where

$$g_i = \frac{\theta_i + \bar{t}_{i.}}{T} \tag{A2.9}$$

and

$$T = \sum_{i=1}^{\eta} (\theta_i + \bar{t}_{i.}) \tag{A2.10}$$

so that

$$\hat{P}(q \geq IB) = \frac{\sum_{i=1}^{\eta} \bar{t}_{i.}}{\sum_{i=1}^{\eta} (\theta_i + \bar{t}_{i.})} \tag{A2.11}$$

Substitution of (A2.7) and (A2.11) into (A2.1) yields

$$\hat{P}(q \geq 15) = \frac{\sum_{i=1}^{\eta} \bar{\tau}_{i.}}{\sum_{i=1}^{\eta} (\theta_i + \bar{t}_{i.})} \tag{A2.12}$$

APPENDIX 3: VARIANCES OF ESTIMATORS IN SIMPLE QUEUING SYSTEM

We derive $\text{var}(\hat{p})$ following Iglehart (1975): Define

$$\underline{z}_i = \underline{\tau}_i - p(\underline{t}_i + \underline{\theta}_i) \tag{A3.1}$$

so that

$$E(\underline{z}_i) = 0 \quad \text{and} \quad \sigma_z^2 = E(z_i^2) \tag{A3.3}$$

We further have

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i \qquad \text{(A3.4)}$$

so that $\bar{z} \sim N(0,\sigma_{\bar{z}})$ for $n \to \infty$. From Iglehart (1975, p. 7) it follows that

$$\frac{\bar{z}}{\sigma_{\bar{z}}} = \frac{\bar{z}}{\sigma_z/\sqrt{n}} = \sqrt{n}\,\frac{\bar{z}}{\sigma_z} \qquad \text{(A3.5)}$$

has a $N(0,1)$ distribution for $n \to \infty$. Substituting (3.1) yields

$$\frac{\bar{z}}{\sigma_{\bar{z}}} = \sqrt{n}\,\frac{(\bar{\tau} - p\bar{T})}{\sigma_z} \simeq \sqrt{n}\,\frac{(\bar{\tau}/\bar{T} - p)}{\sigma_z/E(T)} \qquad \text{(A3.6)}$$

which has an asymptotic standard normal distribution. Consequently

$$\text{var}(\hat{p}) = \text{var}(\bar{\tau}/\bar{T}) \Rightarrow \frac{\sigma_z^2}{E^2(T).n} \qquad (n\to\infty) \qquad \text{(A3.7)}$$

For var(z) we can write

$$\sigma_z^2 = \sigma_\tau^2 - 2p\,\sigma_{\tau,\,t+\theta} + p^2\,\sigma_{t+\theta}^2 \qquad \text{(A3.8)}$$

where $\sigma_{\tau,t+\theta} \equiv \text{cov}(\tau,\ t+\theta)$. Since τ and θ are independent, (3.7) and (3.8) yield

$$\text{var}(\hat{p}) \Rightarrow \frac{\sigma_\tau^2 - 2p\,\sigma_{\tau,t} + p^2\,\sigma_{t+\theta}^2}{n.E^2(t+\theta)} \qquad (n\to\infty) \qquad \text{(A3.9)}$$

When we apply Importance Sampling, 0_i is followed by several timepaths of length t_{ij} . We obtain several τ_{ij}'s. As an estimator we use:

$$\hat{P}_{IS} = \hat{P}_{IS}(q \geq 15) = \frac{\frac{1}{\eta} \sum_{i=1}^{\eta} \bar{\tau}_{i.}}{\frac{1}{\eta} \sum_{i=1}^{\eta} (\theta_i + \bar{t}_{i.})} \qquad \text{(A3.10)}$$

where

$$\bar{\tau}_{i.} = \frac{1}{m} \sum_{j=1}^{m} \tau_{ij} \qquad \text{(A3.11)}$$

and

$$\bar{t}_{i.} = \frac{1}{m} \sum_{j=1}^{m} t_{ij} \qquad \text{(A3.12)}$$

where m denotes the number of replications. Note that (3.10) is biased. To find the variance we proceed analogous to (A3.1): Define

$$\zeta_i = \bar{\tau}_{i.} - p(\bar{t}_{i.} + \theta_i) \qquad \text{(A3.13)}$$

so that

$$\sigma_{\zeta}^2 = \sigma_{\bar{\tau}}^2 - 2p\, \sigma_{\bar{\tau}, \bar{t}+\theta} + p^2 \sigma_{\bar{t}+\theta}^2 \qquad \text{(A3.14)}$$

σ_{ζ}^2 can be related to σ_z^2: We have

(i) $$\sigma_{\bar{\tau}}^2 = \frac{\sigma_{\tau}^2}{m} \qquad \text{(A3.15)}$$

because of the independence of the replecations

(ii) $$\sigma_{\bar{\tau}, \bar{t}+\theta} = \sigma_{\bar{\tau}, \bar{t}} = \frac{\sigma_{\tau, t}}{m} \qquad \text{(A3.16)}$$

where the first equality holds, since τ and θ are independent

$$\text{(iii)} \quad \sigma^2_{\bar{t}+\theta} = \sigma^2_{\bar{t}} + \sigma^2_{\theta} = \frac{\sigma^2_t}{m} + \sigma^2_{\theta} \qquad \text{(A3.17)}$$

where the first equality holds since t and θ are independent. Consequently (A3.14) becomes:

$$\sigma^2_{\zeta} = \frac{1}{m}\,\sigma^2_z + p^2\,\frac{m-1}{m}\,\sigma^2_{\theta} \qquad \text{(A3.18)}$$

For $\text{var}(\hat{P}_{IS})$ we find:

$$\text{var}(\hat{P}_{IS}) \Rightarrow \frac{\frac{1}{m}\,\sigma^2_z + \frac{m-1}{m}\,p^2\sigma^2_{\theta}}{n\,E^2(T)} \qquad (n\to\infty) \qquad \text{(A3.19)}$$

Hence the gross variance reduction is:

$$VR_{gross} \Rightarrow \frac{\sigma^2_z}{\frac{1}{m}\,\sigma^2_z + \frac{m-1}{m}\,p^2\sigma^2_{\theta}} = \frac{m}{1+(m-1)p^2\,\frac{\sigma^2_{\theta}}{\sigma^2_z}} \qquad \text{(A3.20)}$$

$$(n\to\infty)$$

In the main text we derived the net variance reduction:

$$VR_{net} = \frac{m}{\{1+(m-1)P(q\geq 10)\}\{1+(m-1)P^2\frac{\sigma^2_{\theta}}{\sigma^2_{\tau}-2p\sigma_{\tau,t}+p^2\sigma^2_{t+\theta}}\}}$$

(A3.21)

with the constant $\sigma^2_{t+0} = \sigma^2_T$. Since P << 1 we can write

$$VR_{net} \simeq \frac{m}{\{1+(m-1)P(q\geq 10)\}\{1+(m-1)P^2 \frac{\sigma^2_\theta}{\sigma^2_\tau}\}} \qquad (A3.22)$$

As the importance boundary increases, the replications decrease in lengths. Hence σ^2_θ increases and σ^2_τ decreases (compare the geometric distribution). This yields less variance reduction. If then, however, P^2 decreases, the effect increases!

APPENDIX 4: UNBIASEDNESS OF ESTIMATOR (5.2)

Obviously

$$E(\eta_k) = \frac{E(\eta_k)}{E(SSUR_k)} \cdot E(SSUR_k) = P(b|UR) \cdot E(SSUR_k) \qquad (A4.1)$$

where $SSUR_k$ denotes the size of the unimportant region (UR). Hence

$$\frac{1}{SS} E(\eta_k) = P(b|UR) \cdot \frac{E(SSUR_k)}{SS} = P(b|UR) \cdot P(UR) = P(b \cap UR) \qquad (A4.2)$$

We further have

$$E(SSUR_k) = SS - E(\eta_k) \cdot SSR \qquad (A4.3)$$

We assume that η_k and $LR_{ki\,j}$ are independent, which is a realistic assumption. We know that $LR_{ki\,j}$ is independent of the other replications (using different random numbers), say,

$LR_{ki\,j}$, $(j \neq j')$. Finally, we ignore possible dependence between replications in subsequent encounters with an important event within the same subrun k, i.e., $LR_{ki\,j}$ and $LR_{ki'\,j}$ $(i \neq i')$ are assumed to be independent. This assumption is realistic if important regions are "far" apart so that autocorrelations vanish. Hence

$$E\{\frac{1}{m}\sum_{i=1}^{\eta_k}\sum_{j=1}^{m} LR_{ki\,j}\} = E_{\eta}\{E[\sum_{i=1}^{\eta_k}\frac{1}{m}\sum_{j=1}^{m} LR_{ki\,j}\,|\,\eta_k = \eta]\} =$$

$$= E_{\eta}\{\eta_k\, E(LR|\eta_i = \eta)\} = E(\eta).E(LR) \qquad (A4.4)$$

This expression can also be written as

$$E(\eta).E(LR) = E(\eta)\,\frac{E(LR)}{SSR}\,SSR = P(b|IR).SSR.E(\eta) \qquad (A4.5)$$

so that

$$E(\eta)\frac{E(LR)}{SS} = P(b|IR).\,\frac{E(\eta)\;SSR}{SS} = P(b|IR).P(IR) =$$

$$= P(B\cap IR) \qquad (A4.6)$$

Substituting (A4.2) and (A4.6) into (5.2) yields

$$E(\hat{B}) = P(b\cap UR) + P(b\cap IR) \qquad (A4.7)$$

Since UR and IR are "mutual exclusive and exhaustive" we obtain

$$E(\hat{B}) > P(b) \qquad (A4.8)$$

APPENDIX 5: VARIANCE OF $\hat{B}_k$

Applying (5.4) to (5.3) yields

$$\text{var}(b_k) = \underset{\eta}{E}\{\text{var}(\underline{\eta}_k + \sum_{i=1}^{\eta_k} \frac{1}{m} \sum_{j=1}^{m} LR_{ki\,j} \mid \underline{\eta}_k = \eta\} +$$

$$+ \underset{\eta}{\text{var}}\{E(\underline{\eta}_k + \sum_{i=1}^{\eta_k} \frac{1}{m} \sum_{j=1}^{m} LR_{ki\,j} \mid \underline{\eta}_k = \eta\}$$

$$= T_2 + T_1 \qquad \text{(A5.1)}$$

Using the same assumptions as mentioned above (A4.4) we obtain

$$T_2 = \underset{\eta}{E}\{\eta \ \text{var}(\frac{1}{m} \sum_{j=1}^{m} LR_{ki\,j} \mid \eta_k = \eta)\} =$$

$$= \underset{\eta}{E}\{\frac{\eta}{m} \ \text{var}(LR \mid \eta_k = \eta)\} = \frac{1}{m} E(\eta) \ \text{var}(LR) \qquad \text{(A5.2)}$$

and

$$T_2 = \underset{\eta}{\text{var}}\{\eta + \eta \ E(LR_{ki\,j} \mid \eta_k = \eta)\} = \{1+E\ (LR)\}^2 \ \text{var}(\eta)$$

Hence (A5.3)

$$\text{var}(\hat{B}_k) = \frac{\text{var}(L_k)}{SS^2} = \frac{\{1+E(LR)\}^2 \ \text{var}(\eta) + E(\eta) \ \frac{1}{m} \ \text{var}(LR)}{SS^2}$$

(A5.4)